On the meaning of the weighted alternative free-response operating characteristic figure of merit

Dev P. Chakraborty^{a)}

Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania 15668

Xuetong Zhai

Department of Bioengineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15668

(Received 1 December 2015; revised 22 March 2016; accepted for publication 5 April 2016; published 27 April 2016)

Purpose: The free-response receiver operating characteristic (FROC) method is being increasingly used to evaluate observer performance in search tasks. Data analysis requires definition of a figure of merit (FOM) quantifying performance. While a number of FOMs have been proposed, the recommended one, namely, the weighted alternative FROC (wAFROC) FOM, is not well understood. The aim of this work is to clarify the meaning of this FOM by relating it to the empirical area under a proposed wAFROC curve.

Methods: The weighted wAFROC FOM is defined in terms of a quasi-Wilcoxon statistic that involves weights, coding the clinical importance, assigned to each lesion. A new wAFROC curve is proposed, the y -axis of which incorporates the weights, giving more credit for marking clinically important lesions, while the x -axis is identical to that of the AFROC curve. An expression is derived relating the area under the empirical wAFROC curve to the wAFROC FOM. Examples are presented with small numbers of cases showing how AFROC and wAFROC curves are affected by correct and incorrect decisions and how the corresponding FOMs credit or penalize these decisions. The wAFROC, AFROC, and inferred ROC FOMs were applied to three clinical data sets involving multiple reader FROC interpretations in different modalities.

Results: It is shown analytically that the area under the empirical wAFROC curve equals the wAFROC FOM. This theorem is the FROC analog of a well-known theorem developed in 1975 for ROC analysis, which gave meaning to a Wilcoxon statistic based ROC FOM. A similar equivalence applies between the area under the empirical AFROC curve and the AFROC FOM. The examples show explicitly that the wAFROC FOM gives equal importance to all diseased cases, regardless of the number of lesions, a desirable statistical property not shared by the AFROC FOM. Applications to the clinical data sets show that the wAFROC FOM yields results comparable to that using the AFROC FOM.

Conclusions: The equivalence theorem gives meaning to the weighted AFROC FOM, namely, it is identical to the empirical area under weighted AFROC curve. © 2016 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4947125>]

Key words: observer performance, free-response, figure of merit, AFROC, weighted AFROC, Wilcoxon, equivalence theorem

1. INTRODUCTION

Receiver operating characteristic (ROC) methodology^{1,2} is used for objective observer performance assessment of imaging systems. In an ROC study, radiologists rate cases (i.e., patient images) for confidence in presence of disease. The ratings are ordered labels, and without loss of generality, one can assume that the ratings are integers with higher numbers corresponding to greater confidence in presence of disease. For each confidence level, false positive fraction (FPF) and true positive fraction (TPF) are calculated by separately cumulating the fraction of counts for nondiseased and diseased cases, respectively, that occur at or above the selected confidence level. Connecting adjacent operating points (FPF, TPF), including the trivial ones at (0,0) and (1,1), yields the empirical (or trapezoidal) ROC curve, the

area under which can be used as an objective figure of merit (FOM) for quantifying observer performance in an ROC study. The empirical area has been shown by Bamber³ to be equivalent to the 2-sample Mann–Whitney–Wilcoxon, frequently abbreviated to Wilcoxon statistic.^{4,5} The Wilcoxon statistic, which involves comparisons of ratings of all possible pairings of nondiseased and diseased cases, see definition below, is an empirical measure of the probability that diseased cases are rated higher than nondiseased cases.

The free-response ROC (FROC) method^{6–11} is used to evaluate performance of imaging systems in tasks that involve uncertainty about the number and locations of lesions. The radiologist is asked to search the image in order to detect and localize any potential lesions. Typically the radiologist does not know *a priori* how many lesions there might be present in a case, if any, although there may be some prior

knowledge across the population for specific diseases. The radiologists provide locations (marks) and confidence levels (ratings) of suspicious findings whose perceived confidence levels exceed the radiologist's minimum reporting threshold and consequently the FROC data structure consists of a random number (0 or more) of mark-ratings pairs per case. The marks are classified as lesion localizations (LLs) if they are close to true lesions or nonlesion localizations (NLs) otherwise. The FROC curve is defined as the plot of lesion localization fraction (LLF) vs nonlesion localization fraction (NLF), where LLF is the number of LLs with ratings at or above the selected rating divided by the total number of lesions, and NLF is the number of NLs with ratings at or above the selected rating divided by the total number of cases.

Another free-response operating characteristic implied by Bunch *et al.*^{7,12} was subsequently termed the alternative FROC (AFROC).¹⁰ The ordinate of the AFROC is identical to that of the FROC, i.e., LLF. The abscissa, FPF, is defined as follows: (i) for each nondiseased case, the data are used to infer a highest rating ROC false positive (FP) rating or negative infinity if the case has no marks and (ii) the cumulated fraction of nondiseased cases with FP rating at or above the selected confidence level is FPF. The AFROC is defined as the plot of LLF vs FPF. Henceforth, "highest-rating inferred" is implied any time one is dealing with ROC-like quantities derived from FROC data.

A potential issue with the AFROC FOM is that, while it has not been explicitly shown, a diseased case with many lesions can have inordinate influence as compared to a diseased case with one lesion. This is undesirable because patients are generally regarded as equally valid representatives of the population from which they are sampled, and there is no reason to give more importance to some patients over others. A weighted AFROC (wAFROC) FOM defined in terms of a quasi-Wilcoxon statistic has been defined¹¹ that presumably overcomes this problem, but this too has not been explicitly shown. The scheme assigns numerical values (weights) to the lesions on a diseased case, where the weights sum to unity. For example, if there were one lesion its weight would be unity, if there are two lesions the weights could be 0.5 each or 0.9 for the 1st lesion and 0.1 for the 2nd, etc. The weights are intended to code the clinical importance of each lesion, with the observer getting more credit for marking clinically important lesions over less important ones. All published statistical validity and power testing^{11,13} are limited to ROC, AFROC, and other FROC-curve based FOMs. Specifically, the wAFROC FOM has not been similarly tested.

Bamber's 1975 work in ROC analysis is the starting point for all nonparametric methods of analyzing ROC curves, e.g., Refs. 14 and 15. Prior to Bamber's work, one knew how to plot an empirical operating characteristic and how to calculate the Wilcoxon statistic, but their equality had not been analytically proven. This was Bamber's essential contribution. In the absence of this theorem, in the context of ROC analysis, the Wilcoxon statistic would be "just another statistic." The theorem is so important that a major paper appeared in radiology¹⁶ devoted to the equivalence. The title of this paper was "*The meaning and use of the area*

under a receiver operating characteristic (ROC) curve." The equivalence theorem gives meaning to the area under the ROC.

The introduction of the weighted AFROC figure of merit left unresolved issues: (1) while one could calculate it, one did not know how to plot the corresponding weighted AFROC curve (Bamber did not face the analogous problem because how to plot ROC curves was well-known in 1975). (2) Assuming, one can plot the weighted AFROC curve, one still has to analytically prove the equality between the quasi-Wilcoxon statistic and the area under this new curve. Solving these two unresolved issues is the main focus of the current work. A secondary purpose of this paper is to show explicitly, using a simple example, that the AFROC FOM gives unequal importance to abnormal cases with different numbers of lesions and that the weighted AFROC FOM corrects this problem. Finally, since there is limited information currently available on the performance of the wAFROC FOM in modality-comparison studies, the most common application of observer performance methodology, we wished to compare it to AFROC and ROC FOMs using real clinical data sets.

2. METHODS

The ROC curve and the Wilcoxon statistic are defined, along with a statement of equivalence between the empirical area under the ROC and the Wilcoxon statistic. The AFROC curve, the weighted AFROC curve, and the inferred ROC (iROC) curves are defined. A theorem of equivalence between the empirical area under the weighted AFROC and a quasi-Wilcoxon statistic is proved. Examples are given with small numbers of cases, amenable to hand-calculation, of the construction of the AFROC and weighted AFROC curves.

2.A. The ROC curve and the Wilcoxon statistic

Cases are indexed by k_it , where t indicates the disease-status at the case (or patient) level, with $t = 1$ for nondiseased cases and $t = 2$ for diseased cases. The index k_1 ranges from 1 to K_1 for nondiseased cases and k_2 ranges from 1 to K_2 for diseased cases, where K_1 and K_2 are the total number of nondiseased and diseased cases, respectively. Let z_{k_it} represent the z -sample of case k_it . The z -sample is the actual perceived decision variable for the case, which could be a floating point number and is generally unobservable. The ratings, generally chosen to be integers, represent the ordered labels assigned to binned z -samples. Assuming the number of bins is large enough, there is no loss of ordering information in going from z -samples to ratings, and the two terms can be used interchangeably.

For a given reporting threshold ζ and assuming higher values of threshold correspond to greater confidence in presence of disease, FPF(ζ) and TPF(ζ) are defined by

$$\left. \begin{aligned} \text{FPF}(\zeta) &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(z_{k_11} \geq \zeta) \\ \text{TPF}(\zeta) &= \frac{1}{K_2} \sum_{k_2=1}^{K_2} I(z_{k_22} \geq \zeta) \end{aligned} \right\}. \quad (1)$$

Here, $I(x)$ is the indicator function that equals 1 if x is true and 0 otherwise.

The ROC curve is the plot of TPF(ζ) vs FPF(ζ) as the threshold ζ is varied over the range $[-\infty, +\infty]$ plus, if necessary, a straight-line segment connecting the uppermost nontrivial operating point to (1,1). The Wilcoxon statistic, as applied to the two arrays $z_{k_1 1}$ and $z_{k_2 2}$, is defined by

$$W = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \psi(z_{k_1 1}, z_{k_2 2}). \quad (2)$$

In Eq. (2), the kernel function $\psi(x, y)$ is defined by

$$\begin{aligned} \psi(x, y) &= 1 & \text{if } x < y \\ \psi(x, y) &= 0.5 & \text{if } x = y. \\ \psi(x, y) &= 0 & \text{if } x > y \end{aligned} \quad (3)$$

It has been shown that the above Wilcoxon statistic yields the same value as the area under the empirical (or trapezoidal) ROC curve³ (Bamber's theorem).

2.B. The AFROC curve

Free-response decision sites, or simply sites, are indexed by $l_s s$ where s indicates the truth at the site level, with $s = 1$ for noise sites and $s = 2$ for signal sites; $l_1 = 1, 2, \dots$ indexes noise sites, provided the image has at least one noise site, and otherwise $l_1 = \emptyset$, the null set. Likewise, $l_2 = 1, 2, \dots, L_{k_2}$ indexes signal sites, provided the image has at least one signal site, and otherwise $l_2 = \emptyset$. Here, $L_{k_2} > 0$ is the total number of lesions visible on diseased case $k_2 2$. Let $z_{k_t t l_s s}$ represent the z -sample of site $l_s s$ on case $k_t t$. Nondiseased cases with $l_1 = \emptyset$ and unmarked lesions are assigned $-\infty$ z -sample. A site is marked, i.e., recorded as a NL or LL, if its z -sample exceeds the reporting threshold ζ . The highest z -sample ROC FP for nondiseased case $k_1 1$ is

$$FP_{k_1 1} = \max_{l_1} (z_{k_1 1 l_1 1}). \quad (4)$$

Here, $\max_{l_1} (z_{k_1 1 l_1 1})$ is the maximum over all z -samples, labeled by the site index l_1 , occurring on nondiseased case $k_1 1$, or $-\infty$ if $l_1 = \emptyset$. The highest- z -sample FPF(ζ) and the LLF(ζ) are defined by

$$\left. \begin{aligned} FPF(\zeta) &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(\max_{l_1} (z_{k_1 1 l_1 1}) \geq \zeta) \\ LLF(\zeta) &= \frac{1}{\sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} I(z_{k_2 2 l_2 2} \geq \zeta) \end{aligned} \right\}. \quad (5)$$

The AFROC curve is defined as the plot of LLF(ζ) vs FPF(ζ) plus a straight-line segment connecting the uppermost nontrivial operating point to (1,1). A related curve, termed the exponentially transformed FROC (EFROC) curve, has been described¹⁷ that like the AFROC has the advantage of being contained within the unit square.

The AFROC FOM is defined¹⁸ by the quasi-Wilcoxon statistic shown below

$$A_{AFROC} = \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} \psi \left(\max_{l_1} (z_{k_1 1 l_1 1}), z_{k_2 2 l_2 2} \right). \quad (6)$$

It can be seen that this FOM achieves its highest value, unity, if and only if every lesion is rated higher than the maximum rating on every nondiseased case, and its lowest value, zero, if and only if every lesion is rated lower than the maximum rating on every nondiseased case.

2.C. The iROC curve

The highest z -sample true positive (TP) z -sample for diseased case $k_2 2$ is

$$TP_{k_2 2} = \max_{l_1 l_2} (z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2}). \quad (7)$$

Here, the maximum is over all z -samples, NLs, and LLs, labeled by the site indices $l_1 l_2$, occurring on diseased case $k_2 2$, or $-\infty$ if $(l_1 = \emptyset) \& (l_2 = \emptyset)$. The highest- z -sample inferred TPF(ζ) is defined by

$$TPF(\zeta) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} I(\max_{l_1 l_2} (z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2}) \geq \zeta). \quad (8)$$

The iROC curve is defined as the plot of TPF(ζ) vs FPF(ζ) plus a straight-line segment connecting the uppermost nontrivial operating point to (1,1).

2.D. The wAFROC curve

Let $W_{k_2 l_2} \geq 0$ denote the weight of lesion $l_2 2$ in case $k_2 2$ (since weights are only applicable to diseased cases, one can, without ambiguity, drop the case and lesion level truth subscripts). For each diseased case $k_2 2$, the weights are subject to the constraint

$$\sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} = 1. \quad (9)$$

We propose that the weighted lesion localization fraction, wLLF(ζ), be defined by

$$wLLF(\zeta) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(z_{k_2 2 l_2 2} \geq \zeta). \quad (10)$$

The wAFROC curve is the plot of wLLF(ζ) vs FPF(ζ) plus a straight-line segment connecting the uppermost nontrivial operating point to (1,1).

2.E. Examples

This work deals with z -samples which, when realized, are classified as NLs or LLs according to their proximity to real lesions. Demonstration of an operating characteristic from z -sample data involves moving a virtual cutoff and cumulating

TABLE I. Simulated data used in example plots for nondiseased cases yielding the green FP circles in Figs. 1 and 3; * = unmarked; NA = not applicable, i.e., simulator did not realize any z -sample.

Nondiseased case	NLs		FPs		Operating points in Figs. 1 and 3
	$k_t t l_s s$	$z_{k_t t l_s s}$	$k_t t$	$FP_{k_t t}$	
1	1111	NA	11	$-\infty$	*
	1121	NA			
2	2111	0.487	21	0.487	F
	2121	NA			
3	3111	0.738	31	0.738	E
	3121	0.576			
4	4111	-0.305	41	-0.305	H
	4121	NA			

fractions. Since an actual reporting threshold is not applicable, no marking occurs, and it would be more precise to state that these are potential NLs or potential LLs, in the sense that they would be marked if the actual reporting criterion was sufficiently low. In the succeeding description, the adjective “potential” is implied.

The following examples illustrate the construction of AFROC and weighted AFROC curves. The data set consists of $K_1 = 4$ nondiseased cases and $K_2 = 4$ diseased cases. Small numbers of cases were deliberately chosen to make the example easier to follow and permit hand-calculation of the operating points. The first two diseased cases have one lesion each and each of the remaining two have two lesions. The z -samples, generated using an FROC simulator based on the radiological search model,^{18–20} are shown in Table I for nondiseased and Table II for diseased cases, respectively. The details of the simulator are irrelevant to the purpose of these examples.

In Tables I and II, the columns labeled $k_t t l_s s$ list the required case-lesion referencing subscripts, the columns labeled $z_{k_t t l_s s}$ list the corresponding z -samples, when realized and otherwise NA (for not applicable or not realized) is listed. The tables show that the simulator did not realize any z -sample on the 1st nondiseased case, and for the 2nd lesion on the 3rd diseased case (with FROC data, some cases may

have no marks and some lesions may go undetected). Because nondiseased cases have no lesions, all realized z -samples listed in Table I are NLs. In contrast, in Table II, each case can generate NLs and LLs.

Table I illustrates the conversion of the NL z -samples to FP z -samples according to Eq. (4). The 2nd column of Table II lists the number of lesions per diseased case L_{k_2} ; columns 3 and 4 list NL z -samples and columns 5 and 6 list the corresponding LL z -samples. Column 8 illustrates the conversion of the NL and LL z -samples to TP z -samples according to Eq. (7).

2.F. The AFROC curve

In Fig. 1, FPs and LLs, represented by green and red circles, respectively, are shown ordered, from left to right, with higher z -samples to the right (referred to as a “linear” depiction of the data). Each circle is labeled using the z -sample subscripting notation $k_t t l_s s$ introduced previously. For example, the right-most red circle corresponds to the LL z -sample originating from the 1st lesion in the 4th diseased case, i.e., z_{4212} . Consistent with Table I, there are three green circles (FPs) not counting FP_{11} , which occurs at $z = -\infty$ representing the 1st nondiseased case with unrealized z -samples. Likewise, consistent with Table II, there are five red circles (LLs) not counting $z_{3222} = -\infty$ representing the unmarked 2nd lesion on the 3rd diseased case.

Starting from $+\infty$, moving the virtual cutoff continuously to the left generates the AFROC curve, Fig. 2. As each FP is crossed, the operating point moves to the right by

$$\frac{1}{K_1} = 0.25. \quad (11)$$

As each LL is crossed, the operating point moves up by

$$\frac{1}{\sum_{k_2=1}^{K_2} L_{k_2}} = \frac{1}{6}. \quad (12)$$

Since it has one lesion, crossing the z -sample for the 1st case would result in an upward movement of $1/6$, and the same is true for the 2nd case. Since the 3rd case contains two

TABLE II. Simulated data used in example AFROC and wAFROC plots for diseased cases yielding the red LL circles in Figs. 1 and 3; NA = not applicable, i.e., simulator did not realize any z -sample; * = unmarked lesion; note that the z -samples listed under NLs are ignored in AFROC and wAFROC analyses but are used in iROC analysis.

Diseased case	L_{k_2}	NLs		LLs		TPs		Lesion weights $W_{k_2 l_2}$	Operating points in Figs. 1 and 3
		$k_t t l_s s$	$z_{k_t t l_s s}$	$k_t t l_s s$	$z_{k_t t l_s s}$	$k_t t$	$TP_{k_2 t}$		
1	1	1211	1.512	1212	0.852	12	1.512	1	D
		1221	NA	1222	NA			NA	
2	1	2211	NA	2212	-0.215	22	-0.215	1	G
		2221	NA	2222	NA			NA	
3	2	3211	NA	3212	1.588	32	1.588	0.6	C
		3221	NA	3222	NA			0.4	
4	2	4211	NA	4212	2.944	42	2.944	0.4	A
		4221	NA	4222	1.984			0.6	
									B

FIG. 1. A linear depiction of the data in Tables I and II, showing z -samples used for plotting the AFROC; the red circles correspond to potential LLs and the green to potential FPs.

lesions, crossing the corresponding z -samples would result in a net upward movement of the operating point by $1/3$. *This shows explicitly that the traditional method gives greater importance to diseased cases with more lesions.* The jumps from lesions on the same case need not be contiguous; they could be distributed, with intervening jumps from lesions on other cases but eventually the jumps will occur and contribute to the net upward movement. The last columns of Tables I and II label the correspondences of the z -samples to the operating points shown in Fig. 2. Unrealized z -samples, if any, are indicated by an asterisk and assigned the $-\infty$ value. As an example, the jumps due to the two lesions on the 4th case are contiguous: see points A and B in Fig. 2. However, the jumps due to the two lesions on the 3rd case are not contiguous: the 1st lesion gives the point C in Fig. 2, but the unmarked lesion on this case, indicated by the asterisk in Table II, eventually contributes when the operating point moves diagonally from point H to (1,1).

2.G. The wAFROC curve

Figure 3, which is analogous to Fig. 1, is a linear depiction of the data in Tables I and II, but this time the lesion weights, shown in Table II, are incorporated, as indicated by varying

FIG. 2. The empirical AFROC curve for the data shown in Tables I and II. The corresponding depiction is Fig. 2. The operating point labels correspond to the last columns of Tables I and II. The area under the empirical curve is 0.7708.

the size of each red circle to indicate its weight (in Fig. 1, all red circles were of the same size). Also, each lesion is labeled with its rating and weight.

Moving a virtual cutoff continuously to the left generates the wAFROC curve, Fig. 4. The movement of the operating point in response to crossing FPs is the same as before. However, as each LL is crossed, the operating point moves up by an amount that *depends on the specific lesion*

$$\frac{W_{k_2 l_2}}{K_2} = \frac{W_{k_2 l_2}}{4}. \quad (13)$$

Since the first two diseased cases have one lesion each (i.e., unit weights), crossing the corresponding z -samples results in upward jumps of $1/4$, see Fig. 4. According to the weights in Table II, crossing the z -sample of the 1st lesion in the 3rd disease case results in an upward jump of $0.6/4$ and that from the 2nd lesion in same case results in an upward jump of $0.4/4$, for a net upward jump of the 3rd case of $1/4$, the same as for the first two diseased cases. Likewise, crossing the z -samples of the two lesions in the 4th disease case results in upward jump of $0.4/4$, for the 1st lesion and $0.6/4$, for the 2nd lesion, for a net upward jump of $1/4$, which is the same as for the first three diseased cases. *This shows explicitly that the weighting method gives each diseased case the same importance, regardless of the number of lesions in it, a property not shared by the area under the AFROC.*

3. RESULTS

3.A. The empirical area under the wAFROC

The area under the wAFROC curve can be calculated by summing the areas of individual trapezoids defined by drawing vertical lines from each pair of adjacent operating points to the bottom edge of the plot. A sample plot is shown in the Appendix, where the summation over the individual trapezoids is performed proving the following theorem:

The net area under the empirical wAFROC curve equals the previously published expression for the weighted AFROC FOM (Ref. 11) in terms of a quasi-Wilcoxon statistic, namely, the empirical area A_{wAFROC} equals

$$A_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \psi(\max_{l_1}(z_{k_1 l_1 1}), z_{k_2 l_2 2}). \quad (14)$$

While not shown explicitly, using the same methods, it can be proved that that area under the empirical AFROC curve equals the AFROC FOM defined in Eq. (6).

FIG. 3. A linear depiction of the data in Tables I and II, including the weights, showing z -samples used for plotting the weighted AFROC curve; the sizes of the red circles code the lesions weights (the lesion-specific weights are shown below each z -sample).

3.B. Clinical applications

Three clinical FROC multiple-reader-multiple-case (MRMC) data sets were used to compare weighted-AFROC with equally weighted lesions, AFROC, and iROC FOMs. Data set A, acquired to compare image processing algorithms used in digital mammography,²¹ consisted of interpretations by four radiologists of 100 nondiseased and 100 simulated diseased mammograms, the latter with 1–3 simulated microcalcifications per diseased case. The original study had five modalities, but for the purpose of this comparison, only two of them, with maximum FOM difference, were used. Data set B, acquired to compare low-dose CT images, acquired for attenuation correction of PET images, with standard diagnostic quality CT images,²² consisted of interpretations by ten radiographers of single axial slice images of an anthropomorphic phantom with inserted lesions. There were 45 lesion free and 47 lesion-present slices and the maximum number of lesions per slice was three. Data set C, acquired to compare digital mammography with breast tomosynthesis,²³ consisted of interpretations by five radiologists of prospectively acquired cases. There were 96 nondiseased and 89 diseased cases and the maximum number of lesions per diseased case was two.

FIG. 4. The empirical wAFROC curve for the data shown in Tables I and II. The corresponding linear plot is Fig. 3. The operating point labels correspond to the last columns of Tables I and II. The area under the wAFROC is 0.7875.

To give an idea of what the wAFROC curve looks like compared to AFROC curves, we show a few examples. For data set A, Figs. 5(a) and 5(b) show AFROC and wAFROC plots for the first reader: plot (a) is for modality 1 and plot (b) is for modality 2. Figures 6(a) and 6(b) show the corresponding reader-average AFROC and wAFROC plots. The lower performance of modality 2 is apparent from both reader 1 and the averaged reader plots. For this data set, the wAFROC curves lie above the AFROC, but preliminary simulation studies suggest that this may not always be true.

The R-language (Ref. 24) package RJafroc (Ref. 25) was used to perform significance testing. The software is an open-source cross-platform enhanced version of Windows JAFROC software²⁶ that has been used in over 95 published FROC studies.²⁷ The results are shown in Table III. Listed in columns 3 and 4 are the reader-averaged FOMs for each modality, the next columns list the FOM differences (modality 1 minus modality 2) and the corresponding 95% confidence intervals, the F -statistics, the denominator degrees of freedom (the numerator degree of freedom was always one) and the p -values.

RJafroc/JAFROC (collectively referred to as JAFROC software) report the results of an overall F -test of the null hypothesis that all modalities being tested have identical FOMs. The analysis obtains two estimates of variance, one due to the differences between modalities and one due to other causes. If the observed ratio of the first variance to the second variance is large enough, and this is the observed F -statistic, the FOMs are expected to be significantly different. The ratio follows the F -distribution, which is characterized by two quantities called the numerator and denominator degrees of freedom. The degrees of freedom have the interpretation of the amount of information available for the corresponding estimate of variance. In general, if the observed F -statistic is large and the degrees of freedom are large, the study tends to be more significant (smaller p -value) and have greater statistical power (power is the ability of the analysis to discriminate between modalities that are actually different).

For data set A, the results in the table are consistent with the lower performance of modality 2 suggested in Figs. 5 and 6. The analyses also reveal: (1) the iROC FOM is always larger than the wAFROC and AFROC FOMs. This is an expected result: the iROC FOM includes instances where the highest NL rating on a diseased case is larger than the LL rating(s) on the case, as is true for the data shown for the 1st diseased case in Table II, and this increases

FIG. 5. Reader 1 AFROC (solid line) and wAFROC (dotted line) plots for data set A: (a) modality 1 and (b) modality 2.

the corresponding probability of exceeding any NL rating on nondiseased cases. An equivalent explanation is that the localization requirement tends to decrease the corresponding FOM because it rules out instances where the highest rating on a diseased case exceeds that on a nondiseased case but the lesion was incorrectly localized. (2) The AFROC and wAFROC FOMs tend to be comparable.¹⁸ (3) The magnitude of the FOM difference tends to be somewhat larger for the two AFROC-based FOMs than for the iROC FOM (but for data set C, they were comparable). This is because the two AFROC-based FOMs have more “room” to vary, since they can vary over the entire range, [0, 1] while the iROC FOM is restricted to the range [0.5, 1]. One expects the difference to be larger for more difficult tasks, where instances of the lesion being missed and/or incorrectly localized tend to be larger. (4) The observed F -statistic, ddf, and p -value, taken collectively, are consistent with greater statistical power of the two AFROC-based FOMs compared to the iROC FOM. This finding has been confirmed using simulation studies¹³ for the AFROC and iROC FOMs, but this work needs to be extended to the wAFROC FOM.

4. DISCUSSION

The primary aim of this study was to establish the equivalence between the wJAFROC FOM and the area under a newly defined operating characteristic termed the wAFROC curve. The wAFROC curve is identified as the plot of wLLF vs FPF plus a straight-line segment connecting the uppermost nontrivial operating point to (1,1). The Appendix proves the equivalence theorem for the wAFROC: it shows analytically that the area under the defined empirical wAFROC curve is identical to that defined via the quasi-Wilcoxon statistic. Unlike the ROC, the AFROC-based figures of merit do not have simple interpretations. The search paradigm and associated FROC data are indeed more complex than the ROC paradigm, and one should not expect a simple FOM. However, and this was one of the aims of the paper, the detailed explanations should allow one to understand the wAFROC and AFROC FOMs and be satisfied that they meet the criteria of valid FOMs, namely, crediting correct decisions and penalizing incorrect ones.

FIG. 6. Reader averaged AFROC (solid line) and wAFROC (dotted line) plots for data set A: (a) modality 1 and (b) modality 2.

TABLE III. Results of MRMC significance testing using weighted AFROC, AFROC, and highest z -sample inferred ROC FOMs (wAFROC = weighted AFROC, iROC = inferred ROC, F -stat = observed F -statistic, ddf = denominator degrees of freedom, p -value = observed P value, CI = confidence interval).

Data set	FOM	Modality 1	Modality 2	FOM difference (95% CI)	F -stat	ddf	p -value
A	wAFROC	0.769	0.714	0.055 (0.033, 0.077)	24.9	56.4	6.06×10^{-06}
	AFROC	0.758	0.703	0.055 (0.032, 0.078)	21.6	236	5.63×10^{-06}
	iROC	0.851	0.808	0.042 (0.01, 0.075)	15.0	3.4	2.40×10^{-02}
B	wAFROC	0.765	0.840	-0.076 (-0.136, -0.015)	6.03	145	1.52×10^{-02}
	AFROC	0.773	0.841	-0.068 (-0.121, -0.015)	6.40	116	1.27×10^{-02}
	ROC	0.847	0.899	-0.052 (-0.104, -0.001)	4.14	48	4.75×10^{-02}
C	wAFROC	0.812	0.711	0.101 (0.036, 0.166)	10.3	24.7	3.67×10^{-03}
	AFROC	0.809	0.706	0.103 (0.039, 0.167)	10.9	23.1	3.09×10^{-03}
	iROC	0.836	0.736	0.100 (0.031, 0.17)	9.18	17.1	7.52×10^{-03}

A possible concern with the use of the wAFROC FOM is that the weighting may sacrifice statistical power and render the method equivalent to ROC analysis in terms of statistical power. Analysis of three clinical data sets suggests that that may not be the case. Since comparing p -values is not the appropriate way to infer statistical power, there is need for more extensive simulation testing to compare the NH and statistical power behaviors of the wAFROC FOM. This study also shows explicitly that the area under the AFROC gives more importance to diseased cases with more than one lesion as compared to cases with one lesion, and that the area under the weighted AFROC corrects this deficiency. As noted earlier, giving equal importance to each diseased case is a desirable characteristic of valid analysis.

As noted in the Introduction, Bamber's equivalence theorem led to much progress in nonparametric analysis of ROC data. Our proofs of the equivalences between the areas under the AFROC and wAFROC and the corresponding quasi-Wilcoxon statistics are simply mathematical proofs. To realize the full potential of these proofs, similar work like that conducted by DeLong *et al.* is needed for the FROC paradigm. This work is not going to be easy one reason being the relative dearth of researchers working in this area, but it is possible. Indeed, work has been published by Popescu¹⁷ on nonparametric analysis of the EFROC plot which, like the AFROC and wAFROC, is completely contained within the unit square. We believe this work can and should be extended to the wAFROC.

A few comments on why a widely used operating characteristic, namely, the FROC plot,²⁶ is not addressed in detail in the current work. A number of FROC curve-based measures including point measures, such as interpolated LLF at a predefined value of NLF,^{8,28} and partial-area measures,²⁹ such as the area under the FROC to the left of a predefined value of NLF have been defined. The choice of the predefined value is arbitrary and there is no consensus as to where to report performance. In addition, there are fundamental problems with FROC curve-based measures. As noted by Popescu,¹⁷ "unlike the receiver operating characteristic... the FROC curve has an undefined, theoretically infinite, right side limit. Therefore, area under the FROC curves cannot be directly used as an overall performance index." This measure does not qualify as a valid FOM that rewards the observer for

correct decisions and penalizes for incorrect ones. In screening mammography, where the prevalence of breast cancer is about 5 per thousand in the U.S., and assuming a typical expert case-level specificity of 90%, about 90% of interpretations of nondiseased cases by experts have no marks. *In the clinical context, these are perfect decisions but, since the FROC curve is defined by the marks, unmarked cases have no influence on FROC-curve based measure.* Likewise, unmarked lesions, which represent worst possible decisions, do not explicitly influence the FROC curve. These problems do not affect the inferred ROC, AFROC, and wAFROC full area under curve measures.

5. CONCLUSION

The equivalence theorem gives meaning to the weighted AFROC FOM, namely, it is identical to the empirical area under weighted AFROC curve. This work elucidates the behavior of the AFROC-based FOMs and shows that they both behave like valid FOMs.

APPENDIX: PROOF OF THE EQUIVALENCE BETWEEN THE AREA UNDER wAFROC CURVE AND THE DEFINITION OF THE wAFROC FOM

The derivation is necessarily more complex than that shown by Bamber³ in 1975 because FROC is a more complex paradigm than ROC and therefore a more complex notation is needed to accommodate it.

With reference to Fig. 7, the operating point labeled i has coordinates $(FPF_i, wLLF_i)$ given by Eqs. (5) and (10), respectively, reproduced here for convenience

$$FPF_i \equiv FPF(\zeta_i) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(\max_{l_1} (z_{k_1 l_1}) \geq \zeta_i),$$

$$wLLF_i \equiv wLLF(\zeta_i) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} w_{k_2 l_2} I(z_{k_2 l_2} \geq \zeta_i). \quad (A1)$$

The area A_i of the leftmost shaded trapezoid in Fig. 7 is

$$A_i = \frac{(FPF_i - FPF_{i+1})(wLLF_i + wLLF_{i+1})}{2}.$$

Then,

$$A_i = \frac{(FPF_i - FPF_{i+1})}{K_2} \frac{1}{2} \left[\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(z_{k_2 2 l_2} \geq \zeta_i) + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(z_{k_2 2 l_2} \geq \zeta_{i+1}) \right].$$

Expanding the first term inside the square bracket

$$A_i = \frac{(FPF_i - FPF_{i+1})}{K_2} \frac{1}{2} \left[\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(\zeta_i \leq z_{k_2 2 l_2} < \zeta_{i+1}) + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(z_{k_2 2 l_2} \geq \zeta_{i+1}) + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(z_{k_2 2 l_2} \geq \zeta_{i+1}) \right].$$

The last two terms are equal, therefore,

$$A_i = \frac{(FPF_i - FPF_{i+1})}{K_2} \left[\frac{1}{2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(\zeta_i \leq z_{k_2 2 l_2} < \zeta_{i+1}) + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} I(z_{k_2 2 l_2} \geq \zeta_{i+1}) \right]. \quad (A2)$$

The final steps of the proof require that the z -samples be converted to integer ratings, which can be done without loss of ordering information if the number of bins is sufficiently large (Bamber's proof³ is likewise limited to integer ratings). Let $r_{k_1 t l_1 s}$ denote the integer rating of mark $k_1 t l_1 s$, which implies that marks with z -samples satisfying $\zeta_i \leq z_{k_1 t l_1 s} < \zeta_{i+1}$, $i = 0, 1, \dots, R$ are rated $r_{k_1 t l_1 s} = i$ (binning rule). The zero rating corresponds to unmarked regions. The dummy cutoffs ζ_0 and ζ_{R+1} are defined as $-\infty$ and $+\infty$, respectively. From Eq. (A1), it follows that

$$\begin{aligned} FPF_i - FPF_{i+1} &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} [I(\max_{l_1}(z_{k_1 1 l_1}) \geq \zeta_i) \\ &\quad - I(\max_{l_1}(z_{k_1 1 l_1}) \geq \zeta_{i+1})] \\ &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(\zeta_i \leq \max_{l_1}(z_{k_1 1 l_1}) < \zeta_{i+1}). \end{aligned}$$

FIG. 7. An example wAFROC curve; from left to right, the two shaded areas correspond to A_i and A_0 , respectively, defined in text below.

Because of the binning rule, $I(\zeta_i \leq \max_{l_1}(z_{k_1 1 l_1}) < \zeta_{i+1})$ can be replaced by $I(\max_{l_1}(r_{k_1 1 l_1}) = i)$, $I(\zeta_i \leq z_{k_2 2 l_2} < \zeta_{i+1})$ can be replaced by $I(r_{k_2 2 l_2} = i)$, and $I(z_{k_2 2 l_2} \geq \zeta_{i+1})$ can be replaced by $I(r_{k_2 2 l_2} > i)$. Then, Eq. (A2) can then be rewritten as

$$\begin{aligned} A_i &= \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \\ &\quad \times \left[\frac{1}{2} I(\max_{l_1}(r_{k_1 1 l_1}) = i) I(r_{k_2 2 l_2} = i) \right. \\ &\quad \left. + I(\max_{l_1}(r_{k_1 1 l_1}) = i) I(r_{k_2 2 l_2} > i) \right]. \quad (A3) \end{aligned}$$

Summing over all values of i , one gets for the total area under the empirical wAFROC curve

$$\begin{aligned} A_{wAFROC} &= \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \\ &\quad \times \left[\frac{1}{2} I(r_{k_2 2 l_2} = \max_{l_1}(r_{k_1 1 l_1})) \right. \\ &\quad \left. + I(r_{k_2 2 l_2} > \max_{l_1}(r_{k_1 1 l_1})) \right]. \quad (A4) \end{aligned}$$

Explanation: Eq. (A4) follows from the property of the indicator function, which constrains i in the indicator functions inside the square bracket in Eq. (A3) to $\max_{l_1}(r_{k_1 1 l_1})$, where the functions are unity and otherwise they are zero. Using the definition of the Wilcoxon kernel function, Eq. (3), it follows that

$$A_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \psi(\max_{l_1}(r_{k_1 1 l_1}), r_{k_2 2 l_2}). \quad (A5)$$

□

- ^{a)}Author to whom correspondence should be addressed. Electronic mail: dpc10ster@gmail.com
- ¹C. E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.* **21**(9), 720–733 (1986).
  - ²C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**(4), 283–298 (1978).
  - ³D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *J. Math. Psychol.* **12**(4), 387–415 (1975).
  - ⁴H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Stat.* **18**, 50–60 (1947).
  - ⁵F. Wilcoxon, "Individual comparison by ranking methods," *Biometrics* **1**, 80–83 (1945).
  - ⁶J. P. Egan, G. Z. Greenburg, and A. I. Schulman, "Operating characteristics, signal detectability and the method of free response," *J. Acoust. Soc. Am.* **33**, 993–1007 (1961).
  - ⁷P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free-response approach to the measurement and characterization of radiographic-observer performance," *J. Appl. Photogr. Eng.* **4**, 166–171 (1978).
  - ⁸D. P. Chakraborty, E. S. Breatnach, M. V. Yester, B. Soto, G. T. Barnes, and R. G. Fraser, "Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules," *Radiology* **158**, 35–39 (1986).
  - ⁹D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," *Med. Phys.* **16**(4), 561–568 (1989).
  - ¹⁰D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," *Radiology* **174**, 873–881 (1990).
  - ¹¹D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis and validation," *Med. Phys.* **31**(8), 2313–2330 (2004).
  - ¹²P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free-response approach to the measurement and characterization of radiographic-observer performance," *Proc. SPIE* **127**, 124–135 (1977).
  - ¹³D. P. Chakraborty, "Validation and statistical power comparison of methods for analyzing free-response observer performance studies," *Acad. Radiol.* **15**(12), 1554–1566 (2008).
  - ¹⁴J. A. Hanley and K. O. Hajian-Tilaki, "Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update," *Acad. Radiol.* **4**(1), 49–58 (1997).
  - ¹⁵E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics* **44**, 837–845 (1988).
  - ¹⁶J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**(1), 29–36 (1982).
  - ¹⁷L. M. Popescu, "Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve," *Med. Phys.* **38**(10), 5690–5702 (2011).
  - ¹⁸D. P. Chakraborty and H. J. Yoon, "JAFROC analysis revisited: Figure-of-merit considerations for human observer studies," *Proc. SPIE* **7263**, 72630T (2009).
  - ¹⁹D. P. Chakraborty, "ROC curves predicted by a model of visual search," *Phys. Med. Biol.* **51**, 3463–3482 (2006).
  - ²⁰D. P. Chakraborty, "A search model and figure of merit for observer data acquired according to the free-response paradigm," *Phys. Med. Biol.* **51**, 3449–3462 (2006).
  - ²¹F. Zanca, D. P. Chakraborty, C. Van Ongeval, J. Jacobs, F. Claus, G. Marchal, and H. Bosmans, "An improved method for simulating microcalcifications in digital mammograms," *Med. Phys.* **35**(9), 4012–4018 (2008).
  - ²²J. D. Thompson, P. Hogg, D. J. Manning, K. Szczepura, and D. P. Chakraborty, "A free-response evaluation determining value in the computed tomography attenuation correction image for revealing pulmonary incidental findings: A phantom study," *Acad. Radiol.* **21**(4), 538–545 (2014).
  - ²³I. Andersson, D. M. Ikeda, S. Zackrisson, M. Ruschin, T. Svahn, P. Timberg, and A. Tingberg, "Breast tomosynthesis and digital mammography: A comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings," *Eur. Radiol.* **18**(12), 2817–2825 (2008).
  - ²⁴*R: A Language and Environment for Statistical Computing*, edited by R-Development-Core-Team (R Foundation for Statistical Computing, Vienna, Austria, 2011).
  - ²⁵X. Zhai and D. P. Chakraborty, RJafroc: Analysis of Data Acquired Using the Receiver Operating Characteristic Paradigm and Its Extensions: R package version 0.1.1, 2015, available at <https://cran.r-project.org/web/packages/RJafroc/index.html>.
  - ²⁶D. P. Chakraborty, "A brief history of free-response receiver operating characteristic paradigm data analysis," *Acad. Radiol.* **20**(7), 915–919 (2013).
  - ²⁷The publication list is viewable at <http://www.devchakraborty.com>.
  - ²⁸L. T. Niklason, N. M. Hickey, D. P. Chakraborty, E. A. Sabbagh, M. V. Yester, R. G. Fraser, and G. T. Barnes, "Simulated pulmonary nodules: Detection with dual-energy digital versus conventional radiography," *Radiology* **160**, 589–593 (1986).
  - ²⁹F. W. Samuelson, N. Petrick, and S. Paquerault, "Advantages and examples of resampling for CAD evaluation," in *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (IEEE, Arlington, VA, 2007), pp. 492–495.